CONCEPTION ET EXPLOITATION D'UNE BASE DE MÉTADONNÉES DE TRAITEMENTS INFORMATIQUES, REPRÉSENTATION OPÉRATIONNELLE DES CONNAISSANCES DE L'EXPERT

par Yann Abd-el-Kader

Introduction

Rechercher, connaître et utiliser les traitements informatiques du domaine géographique n'est pas toujours tâche facile. L'aide aux utilisateurs et développeurs de l'IGN est le besoin à l'origine de notre travail. De nombreuses connaissances, parfois tacites, manquent au novice confronté à différents types de traitements : les SIG possèdent leurs propres formats de données, interface et langage de programmation associés ; les librairies et programmes développés au sein de l'IGN requièrent des compétences spécifiques en programmation, en gestion des bases de données, en cartographie, en traitement d'images, etc. ; les traitements sous forme de services Web demandent, eux, des connaissances spécifiques sur les protocoles de communication. Plusieurs types de documentation existent (manuels, documentations API, forums, etc.), mais leur dispersion, l'hétérogénéité de leur format et l'absence d'un système d'indexation commun (i.e. de l'utilisation d'un vocabulaire contrôlé commun) ne permettent pas de répondre aux besoins d'information identifiés que de façon imparfaite.

Face à ce constat, l'idée de départ de notre travail était de créer une base de métadonnées puis de construire un système permettant la recherche, la consultation et l'enrichissement. L'analyse des besoins a cependant montré que, pour répondre à certaines requêtes de l'utilisateur, une simple base de métadonnées, dont les informations explicitement présentes sont en nombre nécessairement limité, ne pouvait suffire. Il fallait donc mettre en place des mécanismes de dérivation de l'information s'appuyant sur une **représentation opérationnelle des connaissances d'expert**. En particulier, notre ambition était de fournir des modes d'emploi adaptés au contexte d'utilisation (caractéristiques des données, environnement logiciel, connaissances de l'utilisateur).

Nous avons décidé de suivre une double approche : **documentaire**, et orientée **représentation des connaissances**. En effet, notre ambition était de construire, d'une part, un **Système d'Information (**SI) dans lequel la forme structurée des métadonnées, conformes à notre modèle, rende aisé le développement de l'application Web présentée à l'utilisateur, d'autre part, un **Système à Base de Connaissances** (SBC) doté des capacités d'inférences qui nous permettent de simuler une partie du raisonnement de l'expert.

Les principaux résultats obtenus peuvent se résumer en trois points :

- définition d'un modèle de métadonnées ;
- développement d'un SI ;
- développement d'un SBC.

Définition d'un modèle de métadonnées

L'état de l'art dressé au début de notre travail n'a permis de déceler aucun modèle de description des traitements qui réponde pleinement à nos attentes. Nous avons donc défini notre propre modèle, en nous inspirant, notamment, de l'ontologie OWL-S dédiée aux services Web. Organisé selon cinq facettes de description, notre modèle est générique ; il s'applique a priori aux traitements informatiques de n'importe quel domaine. Il prend cependant en compte des aspects spécifiquement adaptés au domaine géographique, tels que la description fine des propriétés des données avant et après traitements et le recours aux illustrations cartographiques. De plus, si la partie grounding d'OWL-S décrit la façon d'accéder à des services Web, la partie mode d'emploi a vocation à décrire la façon d'accéder aux traitements informatiques en général.

Développement d'un SI

Nous avons implémenté notre modèle en XML Schema, créé une base de méta-données XML et développé une application Web qui permet la recherche, la consultation et la saisie des métadonnées. Plusieurs caractéristiques notables mises en œuvre peuvent être relevées. Elles correspondent aux objectifs identifiés lors de l'analyse des besoins et, parfois, tendent à dépasser certaines des limitations qui affectent les documentations classiques. Notamment, nous avons tenté de permettre une description progressive des modes d'emploi, les plus spécifiques héritant des concepts et pré-requis des plus génériques.

Cherchant à permettre l'expression de connaissances générales, nous avons introduit dans notre modèle la notion de famille de traitement. L'expert humain recourt à des exemples pour clarifier ses explications ; l'association de prototypes aux familles de traitements pourra contribuer à rendre plus parlantes nos descriptions. Les illustrations au moyen d'échantillons de données, dont l'intérêt a été montré dans le contexte de la généralisation cartographique par le travail au laboratoire COGIT de F. Hubert (Hubert 2003), ont été intégrées à nos descriptions. Notre application en permet la visualisation et l'acquisition aux formats « image » courants mais aussi au format « vecteur shape ». L'acquisition automatique d'une partie des descriptions est possible. Pour cela nous avons développé un doclet et des programmes basés sur des expressions régulières.

Développement d'un SBC

La recherche de traitement et l'adaptation des modes d'emploi au contexte d'utilisation nécessitent de simuler le raisonnement de l'expert. Les connaissances de ce dernier sont représentées de façon opérationnelle grâce à deux sous-ensembles de la logique du premier ordre : les logiques de description pour les ontologies et les clauses de Horn pour les règles avec variables. Les langages d'implémentation choisis sont ceux du Web sémantique : RDF, OWL et SWRL. Notre base de métadonnées documentaire, traduite dans ces langages, devient une base de connaissances. Pour effectuer sur celle-ci inférences et requêtes, notre système fait appel à la plateforme Sesame 1.2.1 et à un moteur SWRL développé à l'Université libre de Berlin. En marge de l'application, nous avons également expérimenté Jena 2.2.

Apports et limites de notre travail

La proposition du modèle conceptuel de métadonnées des traitements constitue l'un des principaux apports de notre travail ; les exemples créés et les premières expérimentations menées incitent à une certaine confiance quant à l'adéquation aux besoins d'information sur les traitements dans le contexte de l'IGN. Un autre apport est d'avoir montré l'intérêt des principes de représentation des connaissances que sont les logiques de description (LD) et les règles de production en logique du premier ordre dans le but de la simulation d'une partie du raisonnement de l'expert. Les langages de LD permettent de définir des ontologies. Celles que nous avons proposées spécifiquement pour le domaine géographique (fonctionnalités, types de données, problèmes, etc.) demandent à être validées et enrichies par de véritables experts ; elles ont en fait pour fonction principale d'amorcer le processus de spécification formelle des concepts utilisés pour la description des traitements.

Notre travail présente de fortes similitudes avec le projet du Web sémantique ; la présentation imagée de ce dernier sous forme de layer cake illustre d'ailleurs fort bien la progression de notre démarche, de la définition de métadonnées structurées au contrôle des valeurs des éléments de description, puis à l'exploitation de la sémantique des connaissances représentées. Dès lors, l'adoption des langages du Web sémantique RDF, OWL et SWRL pour la mise en oeuvre de notre SBC était un choix naturel. Des enseignements peuvent être tirés des mises en œuvre effectuées, en particulier pour les limites et difficultés auxquelles nous nous sommes heurtés.

Si l'expressivité des langages OWL et SWRL satisfait bien à nos besoins, les moteurs d'inférence existants, pas encore assez matures, n'ont pas permis d'exploiter pleinement les connaissances de notre SBC. Ce problème sera vraisemblablement résolu dans peu de temps avec l'apparition de nouveaux moteurs. Plus délicate en revanche est la question de l'acquisition des connaissances dans le cadre d'un SBC destiné, comme dans notre cas, à être accessible depuis une application Web ouverte aux utilisateurs ``lambda". En effet, notre modèle de métadonnées fournit le cadre dans lequel les connaissances doivent pouvoir s'exprimer ; or une des difficultés dans la conception d'une interface d'acquisition des règles est d'offrir un moyen simple de désigner les éléments de ce modèle, donc de faire référence à des ressources décrites. Il nous a semblé qu'une notation de type DOM était une bonne solution. La traduction automatique en SWRL

des règles acquises sous cette forme s'est avérée quelque peu problématique, quoique non insurmontable. En raison du grand nombre de variables à manipuler, la saisie manuelle de règles SWRL, via un éditeur comme Protégé 3.1, est apparue lourde pour les exemples de complexité moyenne (comme celui de l'adaptation de mode d'emploi ER 3).

Au-delà du cas particulier des règles, l'automatisation de la conversion entre les versions SI et SBC de notre base de métadonnées constitue une difficulté, sinon une limite, de notre application. Le choix d'une architecture duale, également effectué, par exemple, par R. Troncy dans le contexte de l'indexation des documents audiovisuels (Troncy 2004), était nécessaire pour profiter à la fois des avantages d'un schéma XML documentaire ad hoc et des capacités d'inférences d'OWL. Pour autant, il est clair que notre système n'a pas vocation à permettre la conception d'ontologies pour lesquelles des éditeurs spécialisés et des méthodologies spécifiques devront être employés de façon complémentaire. Par ailleurs, nous avons souligné la difficulté qu'il y a à convertir dans notre format XML la partie assertionnelle des ontologies, exprimée en RDF, en raison des multiples syntaxes du langage. L'emploi d'API indépendantes de ces syntaxes, à envisager, impliquerait une certaine lourdeur dans les développements futurs.

Les expériences réalisées semblent confirmer l'adéquation des langages du Web sémantique aux besoins de notre contexte, mais pointent certains problèmes qui se posent à l'usage. Si nous ne les avons pas tous résolus, les développements réalisés incitent à penser qu'il est possible de le faire. Finalement, en leur état actuel, le modèle de métadonnées défini et l'application développée permettent bien de répondre à la majorité des requêtes simples que l'analyse des besoins a fait apparaître. Il reste toutefois de nombreux cas dans lesquels l'utilisateur ne peut être renseigné de façon satisfaisante, comparé à l'information que pourrait fournir un expert humain employant des connaissances générales tacites difficiles à extraire et à modéliser.

Perspectives

L'aide au paramétrage des traitements, complexe dans le domaine géographique, la simulation partielle de leur comportement à des fins de prédiction ou de démonstration, et l'opérationnalisation de connaissances heuristiques pour mieux orienter l'utilisateur, sont quelques-unes des pistes possibles pour poursuivre sur la voie d'un système d'aide « intelligent ». Les scénarios d'adaptation des modes d'emploi que nous avons mis en œuvre sont simples. Cependant, l'utilisateur est parfois confronté à des situations dans lesquelles l'existence de nombreux choix demanderait l'établissement d'un dialogue avec le système. La contrainte de ne sélectionner que des termes proposés, qui est l'un des principes de base de notre interface utilisateur, pourrait alors montrer des limites. Le recours a des outils de TALN devrait permettre à l'utilisateur de s'exprimer de façon plus naturelle.

Chercher à simuler le raisonnement de l'expert et chercher à concevoir des métadonnées qui aident l'utilisateur à raisonner sont deux objectifs distincts que nous avons conciliés, les règles de notre SBC étant des ressources certes opérationnelles mais aussi consultables. Un système d'aide perfectionné nécessitera probablement de représenter des règles internes au système, non destinées à l'utilisateur. Le modèle des métadonnées présentées à ce dernier n'est pas forcément appelé à évoluer. Ce qui devra être amélioré, ce sont les ontologies dont les concepts et individus servent à valuer les éléments de description des traitements. Des méthodes de normalisation sémantique de ces ontologies devront être appliquées (Bac 2000), des consensus entre experts obtenus. Parallèlement, l'intégration et la fusion avec d'autres ontologies du domaine, en particulier celles qui ne vont pas manquer d'apparaître avec le développement annoncé de nombreux services Web géographique, constituent des perspectives de travail nécessaires et prometteuses.

Notre modèle de métadonnées est générique ; il pourra être utilisé pour décrire des traitements informatiques de domaines autres que géographiques. Les principes, langages et techniques mis en oeuvre pour la conception du SI et du SBC sont également génériques ; ils sont issus du domaine de l'ingénierie des connaissances. Un prolongement de notre travail pourrait consister à effectuer la liaison avec le domaine voisin du génie logiciel, accompagnant ainsi la tendance de l'informatique de fournir à l'utilisateur / développeur une vision des traitements affranchie des considérations d'implémentation. Dans ce but et dans celui plus général de descriptions des traitements informatiques, le développement de métadonnées support de la connaissance est un objectif d'avenir.

Bibliographie

Bachimont B., « Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances », dans J. Charlet, M. Zacklad, G. Kassel et D. Bourigault, *Ingénierie des connaissances : évolutions récentes et nouveaux défis*, Paris, Eyrolles, 2000.

Hubert F., *Modèle de traduction des besoins d'un utilisateur pour la dérivation de données géographiques et leur symbolisation par le Web*, thèse de doctorat informatique, Université de Caen, 2003.

Troncy R., *Formalisation des connaissances documentaires et des connaissances conceptuelles à l'aide d'ontologies : application à la description de documents audiovisuels,* thèse de doctorat d'informatique de l'Université Joseph Fourrier – Grenoble I, 2004.

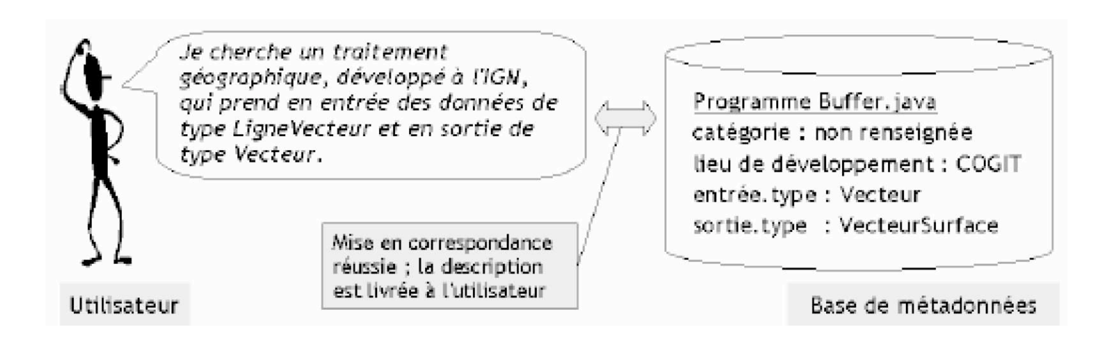


Figure 1 : Recherche de traitements – mise en correspondance entre requête utilisateur et description de traitement

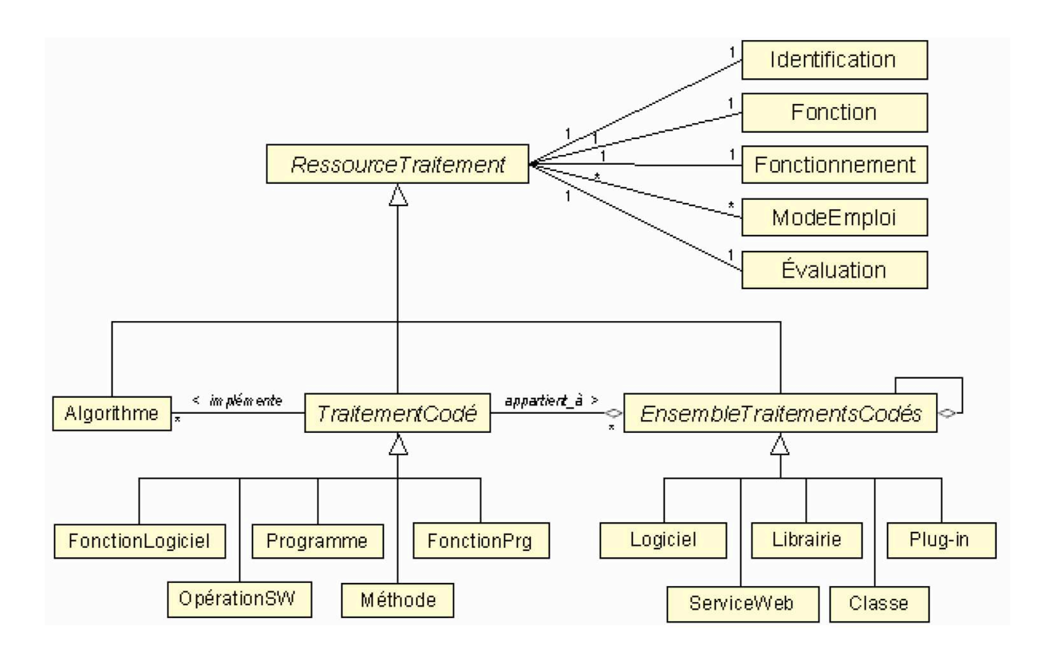


Figure 2 : Classes principales du modèle de métadonnées